Anatolian Journal of Computer Sciences

Volume: 4 No: 1 2019 © Anatolian Science

pp:1-6 ISSN:2548-1304

BİR SOSYAL AĞDAN ALINAN VERİLERİN ANLAMSAL KUTUPLANDIRILMASI

Dilber ÇETİNTAŞ[1] *Taner TUNCER*[2]

*[1]Muş Alparslan Üniversitesi,, Bilgi İşlem Daire Başkanlığı, Muş, Türkiye (d.cetintas@alparslan.edu.tr)

[2]Fırat Üniversitesi, Bilgisayar Mühendisliği Bölümü, Elazığ, Türkiye (ttuncer@firat.edu.tr)

Received: Oct. 31, 2018
Accepted : Dec.19, 2019
Published: Jun. 1, 2019

Özet—İnternetin gelişmesiyle birlikte kullanım oranı her geçen gün artan sosyal ağlar kirli ve düzensiz verilerin bulunduğu ortamlar haline gelmiştir. Verileri düzenleyen ve analiz eden sistemler oluşturabilmek amacıyla bu makalede, twitter sosyal ağından elde edilen Türkçe tweetlerden duygu çıkarımı yapılarak tweetlerinolumlu, olumsuz, nötr olarak sınıflandırılması sunulmuştur. Twitter üzerinden çekilen 20000 verinin sözlük tabanlı doğal dil işleme modeli ile kelimelerin sayısını temel alan özellik vektörleri elde edilmiştir, Elde edilen tweetlerin14471 adedi gönüllü bireyler tarafından etiketlenip %60'ı eğitim %40'ı test verisi olarak kullanılmıştır. Test verisinin sınıflandırılması DVM, Naïve Bayes ve Karar Ağacına göre yapılmıştır. Elde edilen sonuçlara göre en yüksek doğruluk oranı Karar Ağacı ile elde edilmiştir.

Anahtar Kelimeler:Duygu Analizi, Veri Madenciliği, Sınıflandırma Algoritmaları.

1. Giriş

Sosyal ağlar, bireylerininternet iletişim metotları ile iletişime geçmelerini sağlayan, internette toplum yaşamında kendilerini tanımlamasına izin veren, duygu ve düşüncelerinipaylaşabilecekleri, sosyal iletişim kurmalarını sağlayan teknolojilerdir[1].

Yapılan araştırmalarda elde edilen verilere göre Türkiye genelinde internet erişim imkanına sahip hanelerin oranı 2016 yılı Nisan ayında %76,3'dır. İnternet kullanım amaçları dikkate alındığında, 2016 yılının ilk üç ayında internet kullanan bireylerin %82,4'ü sosyal medya üzerinde profil oluşturma, mesaj gönderme veya fotoğraf vb. içerik paylaşırken, bunu %74,5 ile paylaşım sitelerinden video izleme, %69,5 ile online haber, gazete ya da dergi okuma, %65,9 ile sağlıkla ilgili bilgi arama, %65,5 ile mal ve hizmetler hakkında bilgi arama ve %63,7 ile Internet üzerinden müzik dinleme (web radyo) takip etmektedir [2].

Bu istatistiklerde katkı payı yüksek olan 2006 yılında kurulan dünyanın en popüler mikroblog sitelerinden biri olan twitter sosyal ağ sitesidir. Twitter kişilere sunduğu anlık paylaşım yetkisiyle kullanıcıların duyguları hakkında fikir madenciliği yapılarak anormalliklerin tespit edilmesi çalışmalarında yardımcı olmaktadır[3, 4, 5]. Veri madenciliği sonucu oluşan faydalı çıkarımlar, karar vermeyi hızlandırması, kolay erişilebilir olması, oluşabilecek tehditleri öngörüp takip sistemlerinin geliştirilmesi avantajlarını sağlayabilmektedir[6].

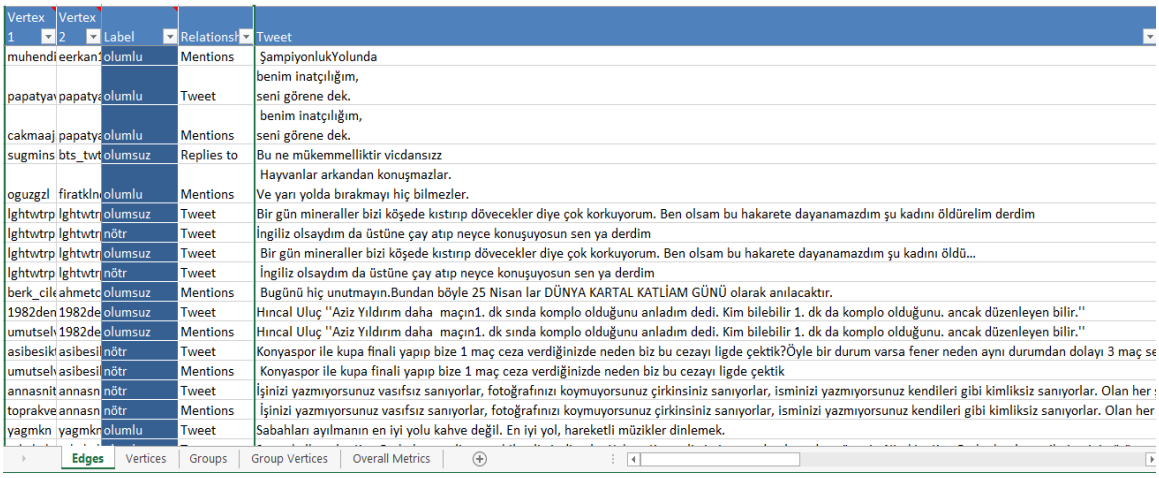
Twitterdan alınan verilere bilgisayar destekli tespit tekniklerini uygulayıp duygu analizi ile anormal kişilikleri en kısa sürede ve doğru olarak tespit edilmesi kötü sonuçları en aza indirgemek açısından önemli bir adım olmaktadır. Verilerin analiz edilmesiyle kirli ve tehlikeli içerikler temizlenebilmekte ve kötü niyetli kişiler ve paylaşımlar sınıflandırılabilmektedir[7].

Bu makaledeki temel amaç paylaşılan içerikleri anlamsal olarak gruplayıp olumsuz olarak

adlandırdığımız toplum ahlakı ile uyuşmayan, rahatsız edici paylaşımları tespit etmektir. Böylece kötü niyetli kullanıcıların tespitinin sağlanması amaçlanmıştır.

2. Veri Alma ve Yapılandırma

Duygu analizinin gerçekleştirilmesinde verilerin alınması ve iyileştirilmesi ilk adımdır. Amaç mevcut verilerin duygu analizi algoritmalarının işleyebileceği formata dönüştürülmesidir. Bu doğrultuda makalede, ilk olarak ağ analizi için NodeXL kullanılmıştır. Temelde bir Excel şablonu olan NodeXL, verilerin alınması, ağın görselleştirilmesi ve analizi için verilerin yapılandırılmasını sağlamaktadır [8]. Verilerin alınması amacıyla parametre olarak girilen (lang:tr) ile Türkçe olarak atılmış tüm tweetler, tweetleri atan düğümler tarih, saat, konumgibi bir dizi veri Excel tablosunda yer alacaktır.Şekil 1 NodeXL ile twitterden alınan tweet'lerin bir bölümünü göstermektedir.

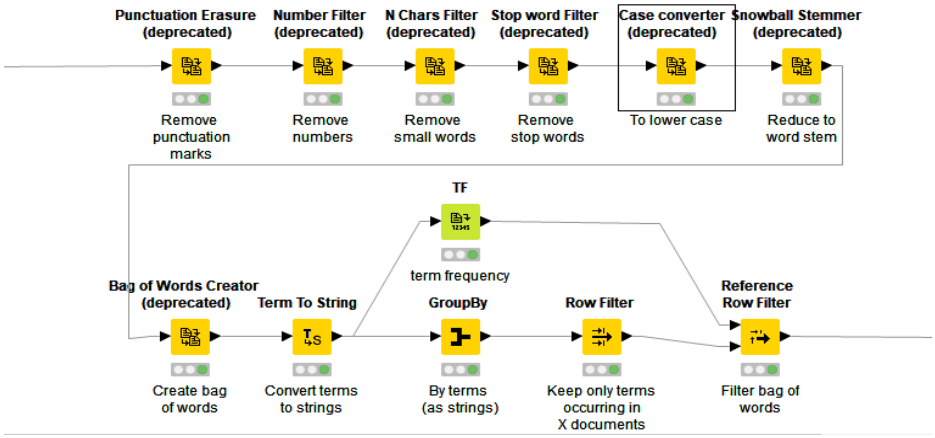


Vertex 1	Vertex 2	Label	Relationship	Tweet
muhendi	eerkan	olumlu	Mentions	ŞampiyonlukYolunda
papatyan	papatya	olumlu	Tweet	benim inatçılığım, seni görene dek.
cakmaaj	papatya	olumlu	Mentions	benim inatçılığım, seni görene dek.
sugmins	bts_twt	olumsuz	Replies to	Bu ne mükemmelliktir vicdansızz
oguzgzl	firatkln	olumlu	Mentions	Hayvanlar arkandan konuşmazlar. Ve yarı yolda bırakmayı hiç bilmezler.
lghtwtrp	lghtwtr	olumsuz	Tweet	Bir gün mineraller bizi köşede kıstırıp dövecekler diye çok korkuyorum. Ben olsam bu hakarete dayanamazdım şu kadını öldürelim derdim
lghtwtrp	lghtwtr	nötr	Tweet	İngiliz olsaydım da üstüne çay atıp neyce konuşuyosun sen ya derdim
lghtwtrp	lghtwtr	olumsuz	Tweet	Bir gün mineraller bizi köşede kıstırıp dövecekler diye çok korkuyorum. Ben olsam bu hakarete dayanamazdım şu kadını öldü...
lghtwtrp	lghtwtr	nötr	Tweet	İngiliz olsaydım da üstüne çay atıp neyce konuşuyosun sen ya derdim
berk_cil	ahmetc	olumsuz	Mentions	Bugünü hiç unutmayın.Bundan böyle 25 Nisan lar DÜNYA KARTAL KATLİAM GÜNÜ olarak anılacaktır.
1982den	1982de	olumsuz	Tweet	Hıncal Uluç "Aziz Yıldırım daha maçın1. dk sında komplo olduğunu anladım dedi. Kim bilebilir 1. dk da komplo olduğunu. ancak düzenleyen bilir."
umutseh	1982de	olumsuz	Mentions	Hıncal Uluç "Aziz Yıldırım daha maçın1. dk sında komplo olduğunu anladım dedi. Kim bilebilir 1. dk da komplo olduğunu. ancak düzenleyen bilir."
asibesik	asibesi	nötr	Tweet	Konyaspor ile kupa finali yapıp bize 1 maç ceza verdiğinizde neden biz bu cezayı ligde çektik?Öyle bir durum varsa fener neden aynı durumdan dolayı 3 maç se
umutseh	asibesi	nötr	Mentions	Konyaspor ile kupa finali yapıp bize 1 maç ceza verdiğinizde neden biz bu cezayı ligde çektik
annasnit	annasn	nötr	Tweet	İşinizi yazmıyorsunuz vasıfsız sanıyorlar, fotoğrafınızı koymuyorsunuz çirkinsiniz sanıyorlar, isminizi yazmıyorsunuz kendileri gibi kimliksiz sanıyorlar. Olan her
toprakve	annasn	nötr	Mentions	İşinizi yazmıyorsunuz vasıfsız sanıyorlar, fotoğrafınızı koymuyorsunuz çirkinsiniz sanıyorlar, isminizi yazmıyorsunuz kendileri gibi kimliksiz sanıyorlar. Olan her
yagmkn	yagmkn	olumlu	Tweet	Sabahları ayılmanın en iyi yolu kahve değil. En iyi yol, hareketli müzikler dinlemek.

Edges | Vertices | Groups | Group Vertices | Overall Metrics

Şekil 1. NodeXL ile Alınan Veriler

Verilerin temizlenmesinden kasıt yanlış ve aşırı uçta bulunan verilerin tweetlerden çıkartılması ve duygu analizini etkileyemeyecek formata getirilmesidir. Bu amaçla Şekil 1' de görüldüğü gibi alınan tweetlerde yer alan RT(Retweet), Hashtag, http ve tweetlerin içerisinde kullanıcılar tarafından yönlendirmek amaçlı verilen linkler Excel'de makro kullanılarak temizlenmiştir.Temizlenen veriler tekrar incelendiğinde farklı lehçe kullanılmasından kaynaklanan ve Türkçe olarak algılanmayan kelimelerin sonucu değiştirmemesi amacıyla manuel olarak düzeltmeler yapılmıştır. Alınan tweetler fikir madenciliği amacıyla kullanılacağı için duygu ifade etmeyen özel isimler, 2 ya da daha az karakterli kelimeler, yanlış yazılımlı ve Türkçe olmayan ifadeler, noktalama işaretleri Şekil2'de gösterilen Knime programı kullanılarak elenmiştir. Bu aşamadan sonra oluşturulan veri seti gönüllü kullanıcılar tarafından olumlu, olumsuz, nötr olarak etiketlenmiştir. Anlamsal karışıklık olabileceği düşünülerek bunu gidermek adına herbir tweet 5 kullanıcıya sunulmuş yüksek sayıda çıkan etiket son karar olarak alınmıştır.Örneğin 3 adet olumsuz 2 adet nötr etiketlenen tweet sonuçta olumsuz olarak etiketlenmiştir.



Şekil 2. Verilerin Knime ile Önişlemesi

Şekil 2'de verilen Knime ön işlemede kullanılan temel işlemlerin kısa açıklamaları aşağıdaki gibidir.

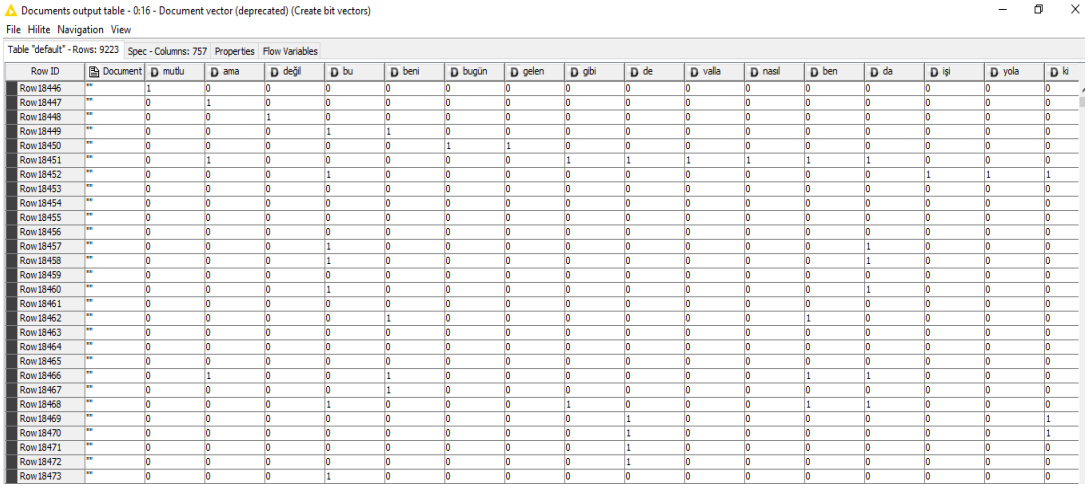
*Snowball:*Türkçe sondan eklemeli bir dil olduğundankelime köküne eklenen yapım eklerinin sonucu etkilememesi için kelime köklerini bulma işlemidir. Türkçe dilinde destek veren bir kütüphane olması sebebiyle tercih edilmiştir.

Stopword İşlemi: Tekrar eden ve tek başına anlam taşımayan kelimelere denir. Bilgiye erişimde stopword listesi, belgeleri bir diğerinden ayırt etme durumuna etkisi olmayan sıklıkla kullanılan kelimeleri içerir. Stopword kelimelerini azaltmak sorgu sürecinin verimini artırmaktadır.

Bag of words: Bu aşamada gruplanan tüm kelimelerin kullanım sıklıkları hesaplanır ve bir havuzda toplanır. Daha sonrasında ise bu kelimelerin değerleri(Word Weighting) hesaplanır. Kelime değeri, bir kelimenin ilgili metin içinde bulunma sıklığıdır [9].

3. Metin işleme

Bilgi çıkarım işlemi, temelde anahtar kelime ve/veya benzerlik tabanlı çıkarımlara dayanmaktadır. Metin dönüşümü varolan metnin kelimeler olarak ele alınıp kelimeleri köklerine ayırma, istenmeyen kelimelerin çıkarılması işlemlerini içerir[4]. Veri analizinde ağırlık verme önemli bir rol oynamaktadır. Çalışmada kullanılan ağırlık verme modeli(TF-Term Frequency) döküman içerisindeki kelimenin tekrar sayısıdır[10]. Kelimelerin ağırlıkları belirlendikten sonra boyutu Bag of Words içerisindeki farklı terimlerin sayısını içeren vektörler oluşturuldu. Şekil 3terimlerden vektörlerin oluşturulmasını göstermektedir.


Documents output table - 0:16 - Document vector (deprecated) (Create bit vectors)

File Hilite Navigation View

Table "default" - Rows: 9223 | Spec - Columns: 757 | Properties | Flow Variables

Row ID	Document	mutlu	ama	değil	bu	beni	bugün	gelen	gibi	de	valla	nasıl	ben	da	işi	yola	ki
Row18446	""	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18447	""	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18448	""	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18449	""	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
Row18450	""	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
Row18451	""	0	1	0	0	0	0	0	1	1	1	1	1	1	0	0	0
Row18452	""	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1
Row18453	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18454	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18455	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18456	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18457	""	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Row18458	""	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Row18459	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18460	""	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
Row18461	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18462	""	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
Row18463	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18464	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18465	""	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Row18466	""	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0
Row18467	""	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Row18468	""	0	0	0	1	0	0	0	1	0	0	0	1	1	0	0	0
Row18469	""	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Row18470	""	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Row18471	""	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Row18472	""	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Row18473	""	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Şekil3. Terimlerin vektörlere dönüştürülmesi

4. Değerlendirme

Literatürde, sınıflandırma başarımlarını belirlemek için Tablo 1.'de gösterilen karışıklık matrisi (Confusion Matrix) kullanılmaktadır. Karışıklık matrisinde satırlar tweetlerin gerçek sınıflarını sütunlar ise sınıflandırma ile tespit edilen tweet sınıfını göstermektedir. Matriste verilen TP,TN, FP, FN değerleri tweet sayısını göstermektedir.

Tablo 1. Karışıklık Matrisi

		Tahmin	
Gerçek		Pozitif	Negatif
	Pozitif	TP	FN
	Negatif	FP	TN

Gerçek pozitif (TP): Doğru pozitif tahmin
Yanlış pozitif (FP): Yanlış pozitif tahmin
Doğru negatif (TN): Doğru negatif tahmin
Yanlış negatif (FN): Yanlış negatif tahmin

Denklem.1 'de verilen hata oranı (ERR), tüm yanlış tahminlerin sayısı, veri kümesinin toplam sayısına bölünerek hesaplanır. En iyi hata oranı 0.0, en kötü ise 1.0'dır.

$$ERR=(FP+FN))/(TP+FN+FP+TN) \quad (1)$$

Denklem.2 'de verilen Doğruluk (ACC), tüm doğru tahminlerin sayısı olarak, veri kümesinin toplam sayısına bölünerek hesaplanır. En iyi doğruluk 1,0, en kötü ise 0,0'dır. Ayrıca 1 - ERR ile hesaplanabilir.

$$ACC=(TP+TN)/(TP+TN+FN+FP) \quad (2)$$

Denklem.3 'de verilen Duyarlılık (SN), doğru pozitif tahminlerin sayısı olarak toplam pozitif sayısına bölünür. Ayrıca Recall (REC) veya gerçek pozitif oran (TPR) olarak da adlandırılır. En iyi duyarlılık 1.0, en kötü ise 0.0'dır.

$$SN=TP/(TP+FN) \quad (3)$$

Denklem.4 'de verilen Özgünlük (SP), toplam negatif sayıyla bölünen doğru negatif tahminlerin sayısı olarak hesaplanır. Aynı zamanda gerçek negatif oranı (TNR) olarak da adlandırılır. En iyi özgüllük 1.0, en kötü ise 0.0'dır.

$$SP=TN/TN+FP \quad (4)$$

Denklem.5 'de verilen Kesinlik (PREC), pozitif tahminlerin toplam sayısına bölünmesiyle doğru pozitif tahminlerin sayısı olarak hesaplanır. Pozitif kestirim değeri (PPV) olarak da adlandırılır. En iyi kesinlik 1.0, en kötü ise 0.0'dır.

$$PREC=TP/(TP+FP) \quad (5)$$

Denklem.6 'da verilen Yanlış pozitif oran (FPR), toplam negatif sayıya bölünen yanlış pozitif tahminlerin sayısı olarak hesaplanır. En iyi yanlış pozitif oran 0.0, en kötü ise 1.0'dır. Ayrıca 1 özgüllük olarak hesaplanabilir.

$$FPR=FP/(TN+FP) \quad (6)$$

Bu çalışmada 21/04/2018 ile 26/04/2018 tarihleri arasında yayınlanmış olan Türkçe tweetler ele alındı. Yakın tarihten geriye doğru alınan tweetlerde 26 Nisan tarihli tweet sayısı daha fazladır. Alınan Tweetlerin 18.000 adedi NodeXL aracılığı ile 2.000 adedi Twitter API ve PHP dili yardımıyla toplam 20000 tweet import edildi. Import edilen bu tweetler 12768 düğüm tarafından atıldığı belirlendi.

Tüm tweet ve tweetlere ait özelliklerin tutulduğu excel dosyası gönüllü kişilere yayılıp olumlu, olumsuz, nötr şeklinde anlamsal kutuplama yapmaları istendi. Gönüllülere sunulan 30 günlük zaman diliminde 14471 tane tweet etiketlenmiş bunlardan 4106 adedi olumsuz, 5244 adedi olumlu, geriye kalan 5121 adedi nötr olarak belirtildi. Tweetlerin alındığı tarihin Türkiyede erken seçim ilanının belirtildiği döneme denk gelmesi sebebiyle talep içeren tweetler(bedelli askerlik gibi …) gönüllüler tarafından nötr olarak etiketlenmiş olup bu durumun nötr sayısını yükselttiği saptandı.

Yinelenen tweetler çıkarıldığında 3592 tweet eksildi. Bunlardan 969 adedi olumsuz, 1350 adedi olumlu, 1276 adedi nötr olduğu saptandı. Bu durum toplum olarak pozitif duyguları daha çok beğenip yaydığımız sonucunu doğurdu.

Sınıflandırma aşamasında ikişerli gruplar halinde alınan(olumlu-olumsuz, olumlu-nötr, olumsuz-nötr) tweetlerin değerlendirilmesi yapılarak herbir etiketin başarıyı etkileme düzeyi araştırıldı. İlk olarak alınan olumlu ve olumsuz veri örneklerinin yer aldığı kümede karar ağacı algoritması ile %82,9 başarı oranı elde edildi. Olumlu verileri doğru tahmin oranı %85,8 olurken olumsuz verilerin doğru tahmin değerinin %79,4'de kaldığı Şekil 4'de gözlemlendi. Bu veri setini Naive Bayes %73,4, DVM algoritması %80,9 doğruluk oranı ile hesapladı.

Accuracy statistics - 0:299 - Scorer

File Hilite Navigation View

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseNe...	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
olumsuz	951	246	1239	205	0.823	0.794	0.823	0.834	0.808	?
olumlu	1239	205	951	246	0.834	0.858	0.834	0.823	0.846	?
Overall	?	?	?	?	?	?	?	?	?	0.829

Şekil 4. Olumlu-Olumsuz En Yüksek Başarı Sonuçları.

Aynı işlemler olumlu-nötr şeklinde alınan veri setine uygulandığında Şekil 5'de verilen karar ağacı algoritması ile %83,5 başarı oranına ulaşıldı. Naive Bayes ile %72,8 DVM ile %80,4 değerleri elde edildi. Bu aşamada nötr ve olumlu değerlerin ayrı ayrı tahmin başarısının birbirine yakın değerler aldığı saptandı.

Accuracy statistics - 0:299 - Scorer

File Hilite Navigation View

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy	Col
nötr	1134	234	1256	237	0.827	0.829	0.827	0.843	0.828	?	?
olumlu	1256	237	1134	234	0.843	0.841	0.843	0.827	0.842	?	?
Overall	?	?	?	?	?	?	?	?	?	0.835	0.67

Şekil 5. Olumlu-Nötr En Yüksek Başarı Sonuçları.

Son olarak ele alınan olumsuz-nötr veri setinde yine karar ağacı ile en yüksek başarı değerine ulaşılmış olup %80 başarı sağlandı. Nötr etiketlerin doğru bulunma oranının olumsuz etiketlerden daha yüksek olduğu Şekil 6'da farkedildi. Bu kategoride Naive Bayes %68,5, DVM %75,5 başarı ile sonuçlandı.

Accuracy statistics - 0:299 - Scorer

File Hilite Navigation View

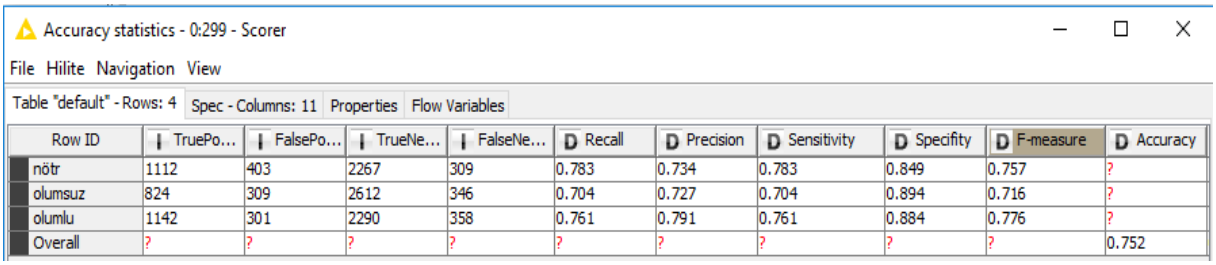
Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseNe...	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
nötr	1127	271	897	235	0.827	0.806	0.827	0.768	0.817	?
olumsuz	897	235	1127	271	0.768	0.792	0.768	0.827	0.78	?
Overall	?	?	?	?	?	?	?	?	?	0.8

Şekil 6. Olumsuz-Nötr En Yüksek Başarı Sonuçları.

Yapılan ikili kombinasyonların sonucunda en yüksek başarının karar ağacı algoritmasıyla en düşük başarının naive bayes algoritmasıyla elde edildiği gözlemlendi. Seçilen veri setinin ve veri setindeki etiket dağılımlarının doğruluk oranını etkilediği saptandı. Olumlu etiket değerine sahip verilerin yer aldığı sınıflandırmalarda olumlu değerlerin doğru tahmin başarısından kaynaklı doğruluk yüzdelerinin arttığı farkedildi. Olumlu etiketlerde yüksek başarının net ifade kullanımından ve toplumun olumlu verileri tekrarlayarak kullanılan kelime ağırlıklarını artırmasından kaynaklandığı; yinelemenin az olması ve mecazi anlatımların, imaların yoğun olarak

kullanılması olumsuz etiketlerde başarının en düşük çıkmasında etkili olduğu sonucuna varıldı.

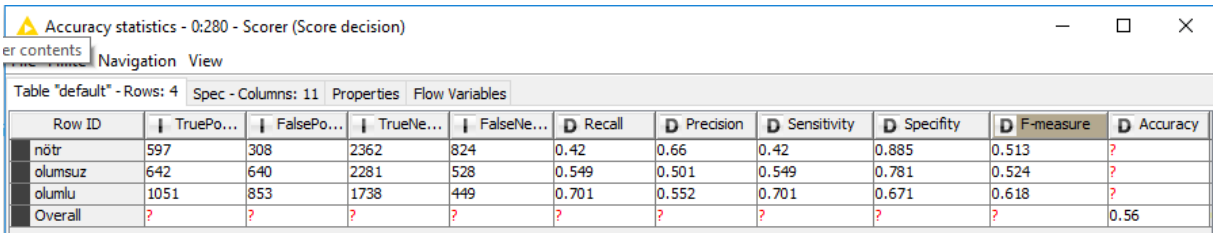

Accuracy statistics - 0:299 - Scorer

File Hilite Navigation View

Table "default" - Rows: 4 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseNe...	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
nötr	1112	403	2267	309	0.783	0.734	0.783	0.849	0.757	?
olumsuz	824	309	2612	346	0.704	0.727	0.704	0.894	0.716	?
olumlu	1142	301	2290	358	0.761	0.791	0.761	0.884	0.776	?
Overall	?	?	?	?	?	?	?	?	?	0.752

Şekil 7. Karar Ağacı Sonuçları.


Accuracy statistics - 0:280 - Scorer (Score decision)

er contents Navigation View

Table "default" - Rows: 4 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseNe...	Recall	Precision	Sensitivity	Specifity	F-measure	Accuracy
nötr	597	308	2362	824	0.42	0.66	0.42	0.885	0.513	?
olumsuz	642	640	2281	528	0.549	0.501	0.549	0.781	0.524	?
olumlu	1051	853	1738	449	0.701	0.552	0.701	0.671	0.618	?
Overall	?	?	?	?	?	?	?	?	?	0.56

Şekil 8. Naive Bayes Sonuçları.

Tüm veri seti sınıflandırma algoritmalarına tabi tutulduğunda en yüksek başarı Şekil 7.' de verildiği gibi Karar Ağacı ile %75,2 oranında, en düşük başarı Şekil 8.' de görüntülenen naive bayes ile %56 oranında elde edildi. Değişken grubunun artmasının başarıyı olumsuz yönde etkilediği farkedildi.

5. Sonuçlar

Sınıflandırma algoritmalarının sonucu işlenecek veri kümesinin ve verilerdeki etiket dağılımının başarı üzerinde ciddi rol oynadığı farkedilmiştir. Olumlu, olumsuz, nötr etiketlerinin dengeli dağılması durumlarında başarı oranının arttığı gözlemlenmiştir. Özellikle nötr etiketinin karar aşamasında etkili olduğu anlaşılmış olup nötr sayının azalmasıyla belirsizliğin azaldığı ve başarının arttığı saptanmıştır. Bu tez çalışmasına katılan gönüllülerin borsa terimlerini içeren ve siyasi görüşlerini bildiren her veriyi nötr olarak etiketlemesi başarı oranını zedelemiştir.

Olumlu etiketine sahip verilerde başarı oranı hep en yüksek olup %80'in altına hiç düşmemiştir. Bu durum geliştirilen sistemin net ifadelerde daha doğru tahminlerde bulunup mecazi anlatım içeren olumsuz ifadelerde aynı başarıyı gösteremediğini kanıtlamıştır.

Başarı durumunu etkileyen bir diğer etmen sadece sözlük tabanlı yöntemin kullanılmış olmasıdır. Örneğin; kullanıcıların İngilizce klavye kullanması durumunda "sınıf" kelimesi "sinif" olarak yazılıp sistem tarafından farklı iki kelime olarak algılanması kelime ağırlıklarını değiştirerek başarı oranını düşürmüştür.

Anlamsal kutuplaşmada sosyal ağ ilişkilerinin etkisi ölçülmeye çalışılmış ağdaki önder kişilerin yaymış olduğu pozitiflik ya da negatiflik alt çizgeleride etkilediği gözlemlenmiştir. Bana arkadaşını söyle sana kim olduğunu söyleyeyim atasözü birkez daha doğrulanmış olup kişi faktörünün ağdaki etiket sayılarını değiştirdiği belirlenmiştir.

Yapılan benzer çalışmalarda hakaret içeren verileri bulma, mutlu ya da üzgün ayrımına gitme gibi daha keskin ayrımlar içeren veya sözlük tabanlı ile n-gram yöntemini birlikte kullanma gibi çalışmalar yapılarak elde edilmiş başarı oranı ile yaklaşık aynı değerlerde başarı elde edilmiştir.

Alınan olumsuz veriler küfür, istismar vs. şeklinde derecelendirilerek güvenlik alanında yapılacak çalışma için önemli rol oynayacaktır.

6. Kaynaklar

[1] https://tr.wikipedia.org/wiki/Sosyal_a%C4%9F, (26/04/2018).

[2] http://www.tuik.gov.tr/PreTablo.do?alt_id=1048, (26/04/2018).

[3] Ö. Nazan, A. Serkan, "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis", Telematics Inf., 35 (1) (2018), pp. 136-147.

[4] Saif H., He Y., Alani H., “Semantic Sentiment Analysis of Twitter”, ISWC 2012. ISWC 2012. Lecture Notes in Computer Science, vol 7649. Springer, Berlin, Heidelberg.
[5] U. Kursuncu, M. Gaur, U. Lokala, K. Thirunarayan, A. Sheth, and I. B. Arpinar, “Predictive Analysis on Twitter: Techniques and Applications,” in Springer-Nature, 2018.
[6] Anurag P. Jain, Vijay D. Katkar, "Sentiments analysis of Twitter data using data mining", 2015 International Conference on Information Processing (ICIP), 2015.
[7] P Selvaperumal and A Suruliandi. 2014. A short message classification algorithm for tweet classification. In Recent Trends in Information Technology (ICRTIT), 2014 International Conference on. IEEE, 1–3.
[8] N.Gürsakal, “Sosyal Ağ Analizi”, Anadolu Üniversitesi Yayınları,Türkiye,2016
[9] U.Çelik, E.Akçetin, M. Gök, “RapidMiner ile Veri Madenciliği”, Pusula Yayınları, 2017.
[10]G.Silahtaroğlu,”Veri Madenciliği Kavram ve Algoritmaları”, Papatya Yayıncılık, Türkiye,2013.